Experimental Design

25

Ethics, Integrity, and the Scientific Method

Jonathan Lewis

Contents

Introduction 460
A Brief Sketch of Experimental Design 461
The Ethical Dimension of Controlled Variation and Randomization 462
The Causal Presuppositions of Experimental Design 466
Reliability, Validity, and Problems of Causal Inference 468
Conclusions 471
References 472

Abstract

Experimental design is one aspect of a scientific method. A well-designed, properly conducted experiment aims to control variables in order to isolate and manipulate causal effects and thereby maximize internal validity, support causal inferences, and guarantee reliable results. Traditionally employed in the natural sciences, experimental design has become an important part of research in the social and behavioral sciences. Experimental methods are also endorsed as the most reliable guides to policy effectiveness. Through a discussion of some of the central concepts associated with experimental design, including controlled variation and randomization, this chapter will provide a summary of key ethical issues that tend to arise in experimental contexts. In addition, by exploring assumptions about the nature of causation and by analyzing features of causal relationships, systems, and inferences in social contexts, this chapter will summarize the ways in which experimental design can undermine the integrity of not only social and behavioral research but policies implemented on the basis of such research.

J. Lewis (✉)
Institute of Ethics, School of Theology, Philosophy and Music, Faculty of Humanities and Social Sciences, Dublin City University, Dublin, Ireland
e-mail: jonathan.lewis@dcu.ie

© Springer Nature Switzerland AG 2020

R. Iphofen (ed.), *Handbook of Research Ethics and Scientific Integrity*,
https://doi.org/10.1007/978-3-030-16759-2_19

Keywords
Experimental design · Randomization · Controlled variation · Deception · Informed consent · Causal relationship · Causal inference · Reliability · Internal validity · External validity · Validity

Introduction

It is possible to distinguish levels of scientific discourse and associated practice (Guala 2009; Arabatzis 2014). The most specific level concerns the precepts of an experiment in a certain discipline. It includes the rules governing the correct use of apparatus and instruments in particular experiments. The next is made up of discourse about experimental design and the practices associated with, for example, laboratory experiments, quasi-experiments, and randomized controlled trials ("RCTs"). At a more abstract level, discussions concern the nature of science in general and, in particular, the nature of scientific theories and the concepts associated with theory appraisal. In recent decades, attention in both science and philosophy has increasingly been paid to discourse about the design and implementation of experimental setups. At the same time, there has been a marked increase in the use of experimental methods in traditionally nonexperimental fields.

With regard to the ethical dimension of experimental design, social and behavioral sciences tend to operate within a framework that was devised primarily with a view to regulating biomedical research. As a result, the methodology literature on the ethics of social and behavioral research often addresses the objectification of research participants, potential harms, purported benefits, coercive and manipulative practices, and issues of privacy and consent. Generally speaking, such concerns do not apply to experimental research in the natural sciences. There will also be issues that arise in one discipline but not another. For example, unlike traditional biomedical experiments, social science research frequently facilitates *interventions* that have winners and losers, create risks for some and not for others, harm or benefit nonparticipants, and operate without the consent of all parties affected by them (Humphreys 2015). Furthermore, not all the ethical issues associated with social and behavioral research result from the employment of experimental methods; many could just as readily occur in the context of nonexperimental studies. Nevertheless, experimental design raises specific ethical problems (though that is not to say that these problems are expressed equally in all branches of empirical research). Indeed, some consider experiments and ethics to be at odds with one another as a result of the tension between the vulnerability of research participants and the interests of pursuing valid and reliable science (Miller and Brody 2003). Others have claimed that experimental design undermines ethical relationships between researchers and participants. For example, experiments not only frequently fail to meet the standards of genuine informed consent, they often involve deception (Sieber 1982).

As a result of various biases, experimental design can also impinge upon the integrity of research, specifically, the *reliability* and *validity* of research. Some sorts

of bias are not particular to experimental design: *value biases* can shape practices at different stages of any research, from the research questions posed to the way they are framed, from decisions about what data to collect to methods for gathering and processing that data, and from the inferences drawn to the reporting and dissemination of results (Douglas 2014). Financial and publication interests, for example, can justify the use of experimental designs in place of observational and more theory-laden approaches (Crasnow 2017). Research funders can impose their values in order to influence experimental design, data interpretation, and the dissemination of results (Wilholt 2009; Brown et al. 2017). Social scientists can act to make the social world more like their models, a "performativity" practice that requires social decision-making and political engagement (Risjord 2014). Value biases can also arise as a result of the tension between non-epistemic values and epistemic values, for example, when considering the possible social consequences of accepting an inference as evidenced when it is not or, conversely, rejecting a claim as invalid when it is, in fact, true. Both non-epistemic and epistemic values can play a part in shaping the reliability and validity of the research.

There are, however, varieties of *inferential bias*, including *selection bias* (errors resulting in overrepresentation of one or more factors in the comparison groups) and failures to take into account *confounders* (omitted variables) (Crasnow 2017), which can arise because experimental methods have been employed. Understanding how these sorts of bias are particular to experimental design requires an engagement with specific theories of causation.

Following a brief introduction to the principal aims and methods of experimental design (section "A Brief Sketch of Experimental Design"), the second section will summarize the key ethical issues that arise when experimenters seek to *control* various variables and *randomly assign* participants to *treatment* and *control* groups (section "The Ethical Dimension of Controlled Variation and Randomization"). The following section will articulate the causal presuppositions of experimental design (section "The Causal Presuppositions of Experimental Design"). Subsequently, the section "Reliability, Validity, and Problems of Causal Inference" will summarize the main problems associated with causal inference in social contexts and thereby illustrate some of the ways in which experimental design can affect the reliability and validity of social and behavioral research as well as any policies implemented on the basis of such research.

A Brief Sketch of Experimental Design

On standard accounts, laboratory experiments aim to isolate purported causes and manipulate causal factors in order to test scientific theories, identify causal relationships, or make particular causal inferences. The laboratory experiment is considered to be an ideal, one that other experimental methods aim to approximate. Indeed, arguments for the value of experimental design tend to invoke the standards of ideal experiments. However, when it comes to investigating social phenomena, laboratory

experiments either cannot be created or they isolate systems to such a degree that inferences within an experiment cannot be extrapolated to *real* social situations (Risjord 2014; Crasnow 2017).

In order to mimic the logic of laboratory experiments in social environments, two experimental methods tend to be employed. The first is quasi-experimentation, whereby two groups are studied in (more or less) "real-world" situations – one group receives the experimental intervention (the "treatment" group), and the other does not (the "control" group). By creating treatment and control groups, the aim is to ensure that circumstances are *controlled* so that the only significant difference between the two groups is some causal factor of interest. In other words, the aim is to rule out all explanations for any observed difference in outcome among those in the treatment group apart from the explanation concerning the average effect of the treatment. By controlling for the other variables in this way, the manipulated variable (X) is considered to be independent of any other possible causes. The problem is that an intervention could produce a correlation between two variables being investigated (X and Y) without the independent variable (X) *causing* the change in the other variable (Y). For example, an intervention could have modified a third variable (Z) *in addition to* the independent variable (X), thereby resulting in a *spurious correlation*.

In order to avoid this kind of correlation, a second experimental method involves the *random assignment* of participants to treatment and control groups. Randomization is meant to ensure that there are no systematic differences between the two groups. The idea is that any potential causes that can affect the outcome of the experiment are evenly distributed. Proponents claim that randomization controls for known, unknown, and unconsidered confounders; it controls for selection bias; it makes the variable on which we are intervening independent of all other variables and thereby allows for conclusions about whether a specific treatment caused the significant difference in outcome between two groups (Conner 1982; Urbach 1985; Papineau 1994; Worrall 2002, 2007; Crasnow 2017). RCTs, which implement randomization in a thoroughgoing way, are considered to be as free from bias as any trial could be outside of the laboratory. Consequently, as Urbach (1985) and Worrall (2002, 2007) suggest, proponents – especially in the biomedical profession – deem RCTs to be necessary for scientific validity or, at the very least, to carry special epistemic weight. Furthermore, and despite the fact that there is little discussion of what justifies these claims, grading schemes for policy effectiveness regularly state that RCTs provide the most reliable scientific evidence, ahead of quasi-experimental research, case studies, and interpretive approaches (Cartwright 2012).

The Ethical Dimension of Controlled Variation and Randomization

The logic of controlled variation is considered to be a hallmark of all forms of experimentation (Guala 2009). In order to make genuine inferences within an experiment, a researcher needs to control the variable that is being manipulated *and* those that are being fixed (Sobel 1996; Goldthorpe 2001; Gangl 2010; Guala

2012). It follows that groups should be situated in conditions whereby extraneous factors are controlled such that the variable of interest can be manipulated in order to observe changes in a second variable. In principle, the employment of a control group facilitates the controlled approach to variable manipulation. However, in order to achieve a sufficient amount of experimental control, researchers in the social and behavioral sciences often employ methods of deception. Even when deception is not used, participants tend not to be fully informed about the research. In addition, for social interventions in particular, researchers regularly do not seek prior consent from their participants. Indeed, informed consent is considered to be impossible in large-scale social experiments (Baele 2013).

A research participant is deceived when any false information is deliberately given or information is deliberately withheld in order to mislead them into believing that something is true when, in fact, it is not (Geller 1982; Sieber 1992; Bordens and Abbott 2013). Consequently, deception is distinct from the usual experimental practice of not fully informing research participants in advance (Hegtvedt 2014). Deception can vary from false information about the main purpose of the study or certain experimental procedures to the presentation of false feedback. It is claimed that deception can undermine a participant's self-confidence or enhance self-doubt, anxiety, and emotional distress and, in extreme cases, can result in the objectification or dehumanization of research participants (Kelman 1982). Even if a certain deception is considered to be fairly innocuous, a standard claim in biomedical ethics is that researchers who either deliberately present false information or deliberately conceal information violate the ethical principle of "respect for autonomy" (Gillon 1994; Beauchamp and Childress 2013). Broadly speaking, by employing deceptive practices, researchers undermine not only a participant's ability to make *genuine* or *authentic* commitments and decisions but a participant's responsibility for those commitments and decisions.

Particularly in experimental research, deceptive practices are methodologically and epistemologically justified. For example, without deception, participant behavior may no longer be "natural" (Clarke 1999). Such *reactive behavior* can undermine the conditions needed for the successful implementation of controlled variation and thereby compromise the reliability of the experiment and the validity of the data. Furthermore, even if a participant is forewarned of the possibility of deception without an account of what it will entail, they may try to determine the nature of the deception and adapt their behavior accordingly (Geller 1982). In order to mitigate these methodological and epistemological concerns, some recommend debriefing sessions for both pretest and study participants. During these sessions, participants can be informed about the deceptive practices employed with the aim of building trust between researchers and participants, demonstrating researcher respect for participants, restoring participants' positive well-being, and removing any self-doubt, anxiety, or emotional distress caused by the experiment (Holmes 1976).

Although deception is considered to be distinct from the practice of not fully informing participants about the research, it is claimed that a deceived participant does not fully understand the nature of the research and, therefore, cannot be fully informed (Clarke 1999). The autonomy of participants can be given a measure of

respect if their consent is sought to use the results of deceptive research after debriefing has taken place. However, this form of post-experimental consent is not an adequate substitute for genuine informed consent. After all, a research participant might not have chosen to participate had they been properly informed about the nature of the experiment and the deceptive practices involved. In such circumstances, it can be argued that not only has informed consent not been given, but a researcher has failed to respect the participant's individual autonomy.

In a laboratory, participants are typically required to give their consent beforehand (though the level of information that is provided can vary from case to case). By contrast, in social experiments, participants tend neither to be informed nor consenting. The reason is that an intervention needs to be perceived by the research participants as naturally occurring (Humphreys 2015). The argument is the same as the one that is used to justify deception; namely, ignorance prevents reactive behavior – Hawthorne effects, John Henry effects, and "looping effects" – that could threaten the scientific outcome by introducing bias (Hacking 1999; Baele 2013). In general, alternative forms of acceptable consent, on the one hand, are challenged on the basis of the methodological and epistemological aims of experimental design. On the other, they are criticized because they do not adequately fulfill the demands of genuine informed consent. This adds weight to the claim that experiments and ethics are fundamentally at odds with one another (Sieber 1982; Miller and Brody 2003).

One of the most frequent ethical concerns raised in connection with controlled variation is the denial of potentially beneficial services for eligible participants (Conner 1982). Moreover, in large-scale social interventions, control-group members (as well as those that do not meet criteria for participation) may actually be worse off as a result of the experiment. Of course, depending on the intervention, control participants may also benefit by not receiving potentially harmful treatments. However, according to one argument, if an experimental intervention is expected to be more beneficial than current social programs, then all individuals of an eligible group should have an equal right to the benefits of the intervention (Conner 1982). Indeed, if a treatment is readily available, then it may be offered to members of the control group after the experiment has concluded. However, if the treatment is beneficial, and if the claims to the treatment are significantly greater for the control group, then, by not ensuring a certain level of well-being for the worse-off group before maximizing the well-being of the treatment group, experimental interventions can violate the prioritarian moral principle (Baele 2013). Furthermore, social interventions tend to operate with relatively scarce resources such that treatments cannot be given to all who stand to benefit from them. It is claimed that the method of randomization, as well as controlling for confounders and selection bias, ensures that all possible targets of a social experiment have an equal chance of being selected to receive the benefits of the intervention (Conner 1982). If a benefit or harm cannot be distributed equally, then randomization – if properly implemented – appears to be a way of guaranteeing that participants are assigned to treatment and control groups in an equitable manner (Lilford and Jackson 1995). In other words, it seems that it can guarantee some sort of fairness.

Randomization, however, is not ethically neutral. Although primarily epistemologically motivated, it carries extraordinary ethical weight (Worrall 2007). By randomly assigning participants to treatments and controls, experiments advantage some people and disadvantage others even though proponents of randomization tend to consider them to be equal (Baele 2013). Conner (1982) argues that such inequities are amplified in laboratory experiments because the participants of the control group are unlikely to receive the standard services available to them in the social world. By contrast, when it comes to RCTs in social settings, control group members are usually not prevented from seeking other available services. Secondly, randomization is not a guarantee of fairness. In order for a potential participant to be treated fairly in a context where a potentially beneficial treatment cannot be distributed equally, their claim to the treatment needs to be taken into account. This is not a right to the treatment but merely a duty to acknowledge the participant's claim to the treatment. As Broome (1984) observes, randomization is a fair option only when the participants potentially affected by an experiment have (more or less) equal claims to the treatment. Consequently, it is not enough for a researcher to advocate randomization in a specific context because they *assume* that all potential participants have equal claims to the treatment; a positive argument is required to show that the claims are *actually* equal (or roughly equal). Randomization, therefore, commits researchers to the judgment that the members of a potential target group have (more or less) equal claims to the treatment. If a beneficial treatment cannot be distributed equally among both treatment and control groups and if certain participants have greater claims to the treatment, then it seems as if the fairest thing to do in that situation would be to give each person a chance proportional to their claim.

In order to justify the use of randomization in controlled trials, researchers in biomedical and public health contexts tend to argue for a necessary state of "clinical equipoise," that is, a state of genuine uncertainty on the part of the expert community regarding the comparative therapeutic merits of each arm in a RCT (Freedman 1987). Some have claimed that there should be a state of "personal equipoise" whereby researchers should be indifferent to the therapeutic value of the experimental and control treatments (Alderson 1996). Others have suggested that both clinical and personal equipoise are necessary for a truly unbiased RCT (Cook and Sheets 2011). The assumption of a state of uncertainty has been identified as the central ethical principle for randomization in human experimentation (Lilford and Jackson 1995). Underlying equipoise is the norm that no patient should be randomized to a treatment known or thought by the expert community to be inferior to the established standard of care (Freedman et al. 1996). This follows from what is perceived to be a clinical researcher's duty of care to their participants (Miller and Weijer 2006), a duty that places severe ethical restrictions on the use of placebo-controlled groups (Fried 1974; Freedman 1987; Miller and Weijer 2006).

It is debatable whether a state of clinical or personal equipoise can be achieved in practice. Indeed, even if a state of equipoise exists prior to the commencement of an RCT, events during a clinical trial, such as the emergence of unexpected adverse events or early signs of efficacy, can quickly disturb clinical and personal equipoise. Furthermore, when it comes to the deployment of experimental methods in

nonclinical settings, specifically, in social contexts, the state of equipoise is either indeterminate or unattainable (Oakley et al. 2003; Hammersley 2008). On the one hand, critics claim that the use of randomization is ethically impermissible because social interventions fail to meet the standards of equipoise in clinical settings. On the other, where a state of equipoise cannot be achieved or when equipoise is dramatically disturbed during the course of a study, randomization cannot escape the ethical questions regarding fairness addressed above. Furthermore, when circumstances make it impossible to distribute a treatment equally to both treatment and control groups, randomization (even when a state of equipoise has been achieved) will violate the duty of fairness if participants' claims to the treatment are not taken into account.

Although the principle of equipoise is adopted to justify randomization, this particular justification is deemed to be necessary when the ethics of research are closely aligned with the duty of care. As we have seen in the context of both controlled variation and randomization, the means to effect experimental design can lead to problems when researchers attempt to uphold that duty. It has been claimed that there is a fundamental conflict between pursuing valid and reliable science and ensuring that no participant is denied a beneficial treatment (Gifford 1986; Miller and Brody 2003). Furthermore, critics argue that this tension cannot be resolved due to a fundamental distinction between the ethics governing experimental research and the ethics of practice-based care (Levine 1979; Churchill 1980; Miller and Brody 2003). Indeed, once we distinguish research ethics from the ethics of care, equipoise can be considered to be no longer relevant to the justification of randomization (Veatch 2007).

The Causal Presuppositions of Experimental Design

An understanding of how experimental design can impinge upon the integrity of research requires a shift from mid-level discussions about experimental design to high-level discourse concerning the causal assumptions held by proponents of experimental design (Guala 2009; Cartwright 2014). From there, it is possible to identify and analyze the problems of causal inference that arise particularly in social contexts and that contribute to the overall validity and reliability of research as well as any policies implemented on the basis of such research.

What is important about experimental practice is not so much observational results and the products of experiments but the design and implementation of experimental setups that reveal or produce causal relationships *in a reliable manner.* In other words, one of main aims of experimental design is to make reliable causal inferences about the effects of some particular causal relationship. In principle, according to the logic of controlled variation, this is done by controlling the variables that are being fixed *and* by isolating and manipulating the purported cause of the variation in the dependent variable. Consequently, causal relationships should be understood in terms of the relations of dependence between variables that remain invariant under interventions. This "interventionist" (also known as a

"manipulationist") theory of causation entails that a fair test of a causal relationship is to apply an intervention on a variable (X) that will change another variable (Y) when, and only when, other causes, preventatives of the effect and other causal relationships involving the effect, are held fixed at some value (Woodward 2008; Cartwright 2014; Kaidesoja 2017). The interventionist view implies that an experiment, when properly implemented, enables researchers to make a clear distinction between spurious correlations and genuine causal relationships by ensuring that the manipulated variable is the only factor that determines the direction of causality.

According to the interventionist theory, two conditions need to be met in order to identify genuine causal relationships (Brady 2011). Firstly, there should be no "interference" across treatment and control groups. In other words, the two groups must be kept separated, isolated, and unable to communicate with each other. The non-interference condition assumes that each supposedly identical treatment really is identical. It also helps to ensure that treatment and control groups are as similar as possible except for the difference in treatment. Brady (2011) claims that if this condition fails, then it will be either difficult to make generalizations about an experiment or impossible to interpret the results. Secondly, treatment and control groups should be, on average, "identical" except for the existence of the putative cause. This "identity" or "unit homogeneity" condition assumes that a variable (X) is independent of any other characteristic that could influence the effect. The two conditions are obviously related; where a control group to interfere with a treatment group or vice versa, the identity condition would no longer hold. Such circumstances could generate (potentially innumerable) unforeseen causal factors that impact upon the causal relationship the researcher is attempting to isolate and manipulate. Consequences of interference include compensatory rivalry between treatment and control groups, resentful demoralization on the part of control groups, attempts by researchers to overcome inequalities between groups, and the diffusion of treatment among members of the control group (Cook and Campbell 1986).

Although unit identity is commonly assumed in laboratory work, it cannot be taken for granted in quasi-experiments and RCTs in social settings (Brady 2011; Risjord 2014). Furthermore, it is impossible to obtain sufficient knowledge about individuals to ensure that the two groups are, on average, identical. If unit homogeneity cannot be guaranteed, then it is a possibility that treatment and control groups are substantially different. It follows that researchers will be required to identify confounders in order to rule out spurious correlations. As we have seen, in order to eliminate the need to identify confounders yet still ensure that an intervening variable (X) is not itself objectively dependent on any other causal factor that could influence the effect on variable (Y), randomization of the study population into experimental and control groups is deemed to be necessary (Papineau 1994; Pearl 2000). There are doubts, however, that randomization can guarantee unit homogeneity to control for confounders and thereby justify causal inferences. While the theoretical underpinnings of randomization support the idea that confounders can be eliminated, it would be a miracle if, in practice, a single random allocation resulted in balanced groups given the innumerable unknown (possible) causes involved (Worrall 2007). Indeed, in any particular case, randomization – even

when properly implemented and even when the two groups seem clearly comparable with respect to all known factors – might result in an unknown or unconsidered distinction between treatment and control groups that plays a significant role in the effect being measured. It is suggested that repeated trials diminish the problem (Binmore 1999; Brady 2011; Crasnow 2017). However, for many social interventions, limited resources make it impossible to facilitate the level of repetition needed to overcome unit imbalance. In addition, repeated draws come with their own risks, primarily an increase in the likelihood of reactive behavior and interference. Consequently, an experimental design with repetition does not automatically guarantee or improve the validity of an experiment (Guala 2009).

Even if randomization cannot, in practice, control for unknown causes, proponents might still claim that it is necessary for ensuring that two groups are comparable with respect to all known factors. However, such balancing can be achieved in a number of ways: by deliberately matching treatment and control groups, by adjusting data ex post, or by checking for "baseline imbalances" (as in clinical trials) and randomizing again when these imbalances are discovered. The point is that randomization is not epistemologically superior to any other method that can be deployed to ensure that there is no positive reason to think that treatment and control groups are unbalanced (Worrall 2007). Of course, proponents might yet claim that randomization is necessary to avoid selection bias. However, Worrall (2007) has shown that, from a theoretical point of view, it is the blinding process, which is effected by randomization, that is ultimately responsible for controlling for selection bias. Again, randomization is not the only means to facilitate this process. Furthermore, it is not clear that, in practice, randomization can fully eliminate selection problems (Crasnow 2017). For all these reasons, it is argued that randomization is neither necessary nor sufficient for guaranteeing genuine causal inferences (Worrall 2002, 2007; Guala 2009).

Reliability, Validity, and Problems of Causal Inference

In the context of experiments involving social phenomena, there are, in general, two types of inference that a researcher can make: firstly, inference from the data/evidence to a cause and, secondly, inference from a particular experiment to other experimental and social contexts (Guala 2012). Given that genuine causal inferences within an experiment demand the ideal conditions of noninterference and unit homogeneity, biases that result from, for example, reactive behaviors, omitted variables, and overrepresentation can affect the strength of the inference drawn from the evidence. Proponents of experimental design in behavioral and social sciences claim that, as a result of randomization, we can be confident that the experimental and control groups are comparable. In addition, by successfully implementing the experimental design, we can isolate different causal factors from possible confounders. Nevertheless, in practice, it is difficult to make a genuine inference that some treatment causes some effect. There is the possibility that interference and reactive behaviors may lead to the reorganization of studied groups in a way that

compensates for the disruption caused by the intervention (Mitchell 2009). Even when aspects of the underlying system have been controlled to the degree that treatment and control groups seem balanced with respect to all known causal factors, there will be many more unconsidered and unknown biases (Worrall 2007). The problem is exacerbated in the case of quasi-experiments; with randomization no longer an option, researchers are forced to make careful determinations about the possible confounders and biases that might be present. The issue is that the reliability and validity of causal inferences based on the data will depend on the ability of the researchers to consider and adjust for these possible confounders and biases, which will depend upon their specific knowledge of the background causal system that gives rise to causal relationships in an experiment (Guala 2005, 2009). Since no real experiment is ideal and since we have no way of knowing how near to the ideal a real experiment is (save for baseline imbalances and other *known* confounders), any particular experiment – whether randomized or not – may mislead about causes (Worrall 2007).

Problems surrounding causal inferences within experiments do not turn solely on the question of whether it is possible, in practice, to control the variables that should be fixed. According to the logic of controlled variation, researchers must also control the variable that is being manipulated. Nevertheless, in the social and behavioral sciences, there can be a number of different variables that cannot be directly manipulated. It can also be extremely complicated – both in theory and in practice – to isolate the intervening variable from the background causal structures of the experiment. Even if it is the case that the variable of interest can be identified, it may not be possible to manipulate it in isolation (Bogen 2002; Woodward 2008). Slight changes in the underlying causal system of an experiment or manipulations of other variables can affect causal processes (Guala 2005), tap into different causal factors (Sullivan 2009), or even subject circumstances to entirely different causal principles (Cartwright 2012). These problems can be dealt with by modifying the ideal design of an experiment accordingly. However, because the circumstances in which social interventions take place are never ideal, changes to the ideal design can trigger a trade-off whereby one problem is solved at the cost of introducing another (Guala 2009).

The legitimacy of causal inferences, as we have seen, depends upon the success of a particular design in controlling for both the variable that is to be manipulated and the variables that are to be held fixed. This issue is typically discussed under the general concept of "validity" (Feest and Steinle 2016). The "internal validity" of an experiment pertains to the inferences within experiments from the data/evidence to the cause. This is contrasted with "external validity," which pertains to the inferences from particular experiments to other contexts. The question of external validity concerns whether the same causal mechanisms operate in other contexts. Some claim that causal inferences made within laboratory experiments are more reliable than those within field experiments precisely because researchers are better equipped to control the relevant variables (Kuorikoski and Marchionni 2014). Others suggest that laboratory conditions simplify the background causal structures of an experiment, thereby making the situation more epistemically manageable (Guala 2005,

2009). Due to the isolation of variables of interest or the simplification of underlying causal systems, target situations are deemed to correspond to the experimental design because the causal relationships involved are believed to be context independent. However, when it comes to the social and behavioral sciences, it cannot be assumed that the causal relationships are context independent to the degree that would allow straightforward inferences from laboratory conditions to the phenomena of interest (Kuorikoski and Marchionni 2014). The methodological or epistemic control exerted in laboratory experiments often results in highly confined and artificial conditions. This is one of the motivations behind field experiments, including quasi-experiments and RCTs, which, according to proponents, have greater external validity precisely because they are conducted in real-world, natural environments over which a researcher has only limited control beyond the intervention (Morton and Williams 2010).

The problem is that greater external validity comes at the cost of internal validity. Due to the fact that social phenomena are situated in highly complex causal systems that give rise to compound causal relationships and mechanisms of causal force, *limited* control over variables can make it difficult to identify genuine causes of social events (Risjord 2014). Nevertheless, proponents claim that the evidence that supports a causal inference in a particular quasi-experiment or RCT is more likely to be generalizable to other real-world contexts. On first appearance, this seems like a fair assumption; if the intervention has, in fact, identified a genuine causal relationship between the manipulated and outcome variables, then presumably the causal factors of interest are context independent such that that relationship can be generalized to other cases. However, this argument depends on important additional assumptions about the similarities between the circumstances under which the experiment is carried out and the circumstances to which the results of the particular experiment are extrapolated. Even if we assume that a genuine causal inference can be drawn within a localized intervention, the flux of the social world may undermine the conclusion that the same causal inference holds beyond the particular case (Crasnow 2017). Consequently, without additional knowledge of the similarities and differences, it may not be justifiable, or even possible, to generalize the causal inferences drawn within a particular experiment to other experimental situations let alone general social contexts. It is widely believed that there is a trade-off between the reliability of our inferences within the confines of the experiment and the reliability of our extrapolations from the experiment. In other words, there is a trade-off between internal and external validity (Guala 2005, 2009; Cartwright 2012; Kuorikoski and Marchionni 2014; Feest and Steinle 2016; Crasnow 2017).

External validity is an important concept in the context of much of today's experimental research. With an ever-increasing concern for "evidence-based policy" and "social impact," public and private research funders tend to favor those projects, disciplines, and research methods that can show us "what works." The effectiveness of a proposed policy is regularly explained and justified on the basis of the reliability of a specific type of research. RCTs and quasi-experiments are typically presented as the most reliable forms of research (Cartwright 2012). As Cartwright (2012) suggests, the effects of an experiment are reliable if they are the results of causes that

happen under the governance of a causal relationship, which, in turn, results from an underlying causal structure. In the case of an experiment that aims to infer a causal relationship between teacher-pupil ratio and pupil achievement, the causal system might include factors such as the frequency of pupil attendance; the ages of the students; whether the teacher is the same in all classroom situations; the fact that schooling is mandatory; the ability, competency, and qualifications of the teacher; the socioeconomic environment in which the school is situated; and so on.

The problem is that the causal relationships involved in social experiments are claimed to be local and fragile (Cartwright 1999, 2007). They are local because they depend on the organization of the underlying causal system and thereby they are deemed to hold only when the "socioeconomic machine" is in place to support them. Policy interventions may involve a different complex of causal factors to that of the initial experiment because of differences in the background causal systems. The causal relationships are fragile because policy interventions made on the basis of a particular causal inference within a specific RCT or quasi-experiment are likely to change the organization of the background system such that the causal relationships no longer hold. Furthermore, by employing randomization in order to insulate an intervention from all known and unknown confounders and biases, even experiments themselves can alter the background causal system that makes a causal relationship possible. As a result, randomization is not a guarantee of external validity.

To warrant the belief that a causal inference can be extrapolated from a particular, well-implemented experiment to other contexts, the proposed techniques of causal inference stress the importance of multiple kinds of evidence and methods (Kuorikoski and Marchionni 2014). For Cartwright (2012), what is required is knowledge of the nature and stability of the causal factors involved and the background causal structures of both the experimental setup and the target setup. Such knowledge, she claims, cannot be underwritten by any particular RCT; it depends on a complicated balance of theory and empirical studies. More importantly, external validity claims require that any additional interventions do not disrupt the causal system that supports the causal relationship identified in the particular RCT or quasi-experiment. Broadly, the problem of extrapolation can be overcome in a similar fashion to the problem of causal inferences within experiments, namely, by way of knowledge of the context and background conditions of the research. It has been suggested that the data provided by experimental design can contribute to external validity only when it is supplemented with analyses of detailed "on-the-ground" evidence generated through multiple, independent nonexperimental methods, including qualitative methods such as case studies and process tracing (Risjord 2014; Crasnow 2017).

Conclusions

This chapter has articulated the main ethical issues associated with experimental design, specifically, those issues that arise when experimental interventions seek to control variables by randomly assigning participants to different groups.

Furthermore, by exploring assumptions about the nature of causation and by analyzing features of causal relationships, this chapter has illustrated some of the ways in which experimental design can undermine the reliability and validity of causal claims thereby affecting the integrity of research and evidence-based policy.

References

Alderson P (1996) Equipoise as a means of managing uncertainty: personal, communal and proxy. J Med Ethics 223:135–139

Arabatzis T (2014) Experiment. In: Curd M, Psillos S (eds) The Routledge companion to philosophy of science, 2nd edn. Routledge, London, pp 191–202

Baele S (2013) The ethics of new development economics: is the experimental approach to development economics morally wrong? J Philos Econ 7(1):2–42

Beauchamp T, Childress J (2013) Principles of biomedical ethics, 7th edn. Oxford University Press, Oxford

Binmore K (1999) Why experiment in economics? Econ J 109(453):16–24

Bogen J (2002) Epistemological custard pies from functional brain imaging. Philos Sci 69(3):59–71

Bordens K, Abbott B (2013) Research and design methods: a process approach. McGraw-Hill, Boston

Brady H (2011) Causation and explanation in social science. In: Goodin R (ed) The Oxford handbook of political science. Oxford University Press, Oxford, pp 1054–1107

Broome J (1984) Selecting people randomly. Ethics 95(1):38–55

Brown A, Mehta T, Allison D (2017) Publication bias in science: what is it, why is it problematic, and how can it be addressed? In: Jamieson K, Kahan D, Scheufele D (eds) The Oxford handbook of the science of science communication. Oxford University Press, Oxford, pp 93–101

Cartwright N (1999) The dappled world: a study of the boundaries of science. Cambridge University Press, Cambridge, UK

Cartwright N (2007) Hunting causes and using them. Cambridge University Press, Cambridge, UK

Cartwright N (2012) RCTs, evidence, and predicting policy effectiveness. In: Kincaid H (ed) The Oxford handbook of philosophy of social science. Oxford University Press, Oxford, UK, pp 298–318

Cartwright N (2014) Causal inference. In: Cartwright N, Montuschi E (eds) Philosophy of social science: a new introduction. Oxford University Press, Oxford, pp 308–337

Churchill L (1980) Physician-investigator/patient-subject: exploring the logic and the tension. J Med Philos 5(3):215–224

Clarke S (1999) Justifying deception in social science research. J Appl Philos 16(2):151–166

Conner R (1982) Random assignment of clients in social experimentation. In: Sieber J (ed) The ethics of social research: surveys and experiments. Springer, New York, pp 57–77

Cook T, Campbell D (1986) The causal assumptions of quasi-experimental practice. Synthese 68(1):141–180

Cook C, Sheets C (2011) Clinical equipoise and personal equipoise: two necessary ingredients for reducing bias in manual therapy trials. J Man Manipulative Ther 19(1):55–57

Crasnow S (2017) Bias in social science experiments. In: McIntyre L, Rosenberg A (eds) The Routledge companion to the philosophy of social science. Routledge, London, pp 191–201

Douglas H (2014) Values in social science. In: Cartwright N, Montuschi E (eds) Philosophy of social science: a new introduction. Oxford University Press, Oxford, pp 162–182

Feest U, Steinle F (2016) Experiment. In: Humphreys P (ed) The Oxford handbook of philosophy of science. Oxford University Press, Oxford, pp 274–295

Freedman B (1987) Equipoise and the ethics of clinical research. N Engl J Med 317(3):141–145

Freedman B, Glass K, Weijer C (1996) Placebo orthodoxy in clinical research II: ethical, legal, and regulatory myths. J Law Med Ethics 24(3):252–259

Fried C (1974) Medical experimentation: personal integrity and social policy. Elsevier, New York
Gangl M (2010) Causal inference in sociological research. Annu Rev Sociol 36:21–47
Geller D (1982) Alternatives to deception: why, what, and how? In: Sieber JE (ed) The ethics of social research: surveys and experiments. Springer, New York, pp 38–55
Gifford F (1986) The conflict between randomized clinical trials and the therapeutic obligation. J Med Philos 11:347–366
Gillon R (1994) Medical ethics: four principles plus attention to scope. Br Med J 309 (6948):184–188
Goldthorpe J (2001) Causation, statistics, and sociology. Eur Sociol Rev 17(1):1–20
Guala F (2005) The methodology of experimental economics. Cambridge University Press, Cambridge
Guala F (2009) Methodological issues in experimental design and interpretation. In: Kincaid H, Ross D (eds) The Oxford handbook of philosophy of economics. Oxford University Press, Oxford, pp 280–305
Guala F (2012) Experimentation in economics. In: Mäki U (ed) Philosophy of economics. Elsevier/North Holland, Oxford, pp 597–640
Hacking I (1999) The social construction of what? Harvard University Press, Cambridge, MA
Hammersley M (2008) Paradigm war revived? On the diagnosis of resistance to randomized controlled trials and systematic review in education. Int J Res Method Educ 31(1):3–10
Hegtvedt K (2014) Ethics and experiments. In: Webster M, Sell J (eds) Laboratory experiments in the social sciences. Academic, London, pp 23–51
Holmes D (1976) 'Debriefing after psychological experiments: I. Effectiveness of postdeception dehoaxing' and 'Debriefing after psychological experiments: II. Effectiveness of post-experimental desensitizing'. Am Psychol 32:858–875
Humphreys M (2015) Reflections on the ethics of social experimentation. J Glob Dev 6(1):87–112
Kaidesoja T (2017) Causal inference and modeling. In: McIntyre L, Rosenberg A (eds) The Routledge companion to philosophy of social science. Routledge, London, pp 202–213
Kelman H (1982) Ethical issues in different social science methods. In: Beauchamp T et al (eds) Ethical issues in social science research. John Hopkins University Press, Baltimore, pp 40–98
Kuorikoski J, Marchionni C (2014) Philosophy of economics. In: French S, Saatsi J (eds) The Bloomsbury companion to the philosophy of science. Bloomsbury, London, pp 314–333
Levine R (1979) Clarifying the concepts of research ethics. Hast Cent Rep 9(3):21–26
Lilford R, Jackson J (1995) Equipoise and the ethics of randomization. J R Soc Med 88(10):552–559
Miller F, Brody H (2003) A critique of clinical equipoise: therapeutic misconception in the ethics of clinical trials. Hast Cent Rep 33(3):19–28
Miller P, Weijer C (2006) Fiduciary obligation in clinical research. J Law Med Ethics 34(2):424–440
Mitchell S (2009) Unsimple truths: science, complexity, and policy. University of Chicago Press, Chicago
Morton R, Williams K (2010) Experimental political science and the study of causality: from nature to the lab. Cambridge University Press, Cambridge, UK
Oakley A et al (2003) Using random allocation to evaluate social interventions: three recent UK examples. Ann Am Acad Pol Soc Sci 589(1):170–189
Papineau D (1994) The virtues of randomization. Br J Philos Sci 45:437–450
Pearl J (2000) Causality-models, reasoning and inference. Cambridge University Press, Cambridge, UK
Risjord M (2014) Philosophy of social science: a contemporary introduction. Routledge, London
Sieber, Joan (1982) Ethical dilemmas in social research. In: Sieber J (ed) The ethics of social research: surveys and experiments. Springer, New York, pp 1–29
Sieber J (1992) Planning ethically responsible research: a guide for students and internal review boards. Sage, Newbury Park
Sobel M (1996) An introduction to causal inference. Sociol Methods Res 24(3):353–379
Sullivan J (2009) The multiplicity of experimental protocols. A challenge to reductionist and non-reductionist models of the unity of neuroscience. Synthese 167:511–539
Urbach P (1985) Randomization and the design of experiments. Philos Sci 52:256–273

Veatch R (2007) The irrelevance of equipoise. J Med Philos 32(2):167–183
Wilholt T (2009) Bias and values in scientific research. Stud Hist Phil Sci 40(1):92–101
Woodward J (2008) Invariance, modularity, and all that. Cartwright on causation. In: Cartwright N et al (eds) Nancy Cartwright's philosophy of science. Routledge, New York, pp 198–237
Worrall J (2002) What evidence in evidence-based medicine? Philos Sci 69(3):316–330
Worrall J (2007) Why there's no cause to randomize. Br J Philos Sci 58(3):451–488